文章编号: 1007-5321(2013)02-0020-04

针对产品命名实体识别的半监督学习方法

黄诗琳， 郑小林， 陈德人
(浙江大学 计算机科学与技术学院，杭州 310027)

摘要: 针对商务信息领域的产品命名实体，研究了产品命名实体各部分的结构特征和相互关系，建立了一个三层的半监督学习框架. 该方法综合利用规则词典和统计的方法，建立一个隐条件随机场模型，可以更充分地利用自举得到数据的隐藏状态. 在数码相机领域进行的实验结果表明，该方法只需要少量的手工标记数据就能较好地识别网页等文本中的产品命名实体.

关 键 词: 产品命名实体识别; 商务信息处理; 自然语言处理
中图分类号: TP181 文献标志码: A

A Semi-Supervised Learning Method for Product Named Entity Recognition

HUANG Shi-lin， ZHENG Xiao-lin， CHEN De-ren
(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: A semi-supervised approach based on a three-level framework for product named entity recognition is presented. The structure features and relationships among different parts of product named entities are studied, and a combined method is applied. A hidden conditional random field model is built so as to utilize the hidden status of learned samples. The labels failed to be learned by the bootstrapping algorithm is considered as hidden statuses. Experiment in digital camera area shows that, with only a few manually labeled data, this method could recognize product named entities from text contents of web pages very well.

Key words: product named entity recognition; business information processing; natural language processing

随着信息技术的发展，市场进入高度信息化时代. 在商务的高度智能化环境下，大量的商务信息以电子文档的形式呈现. 为了给用户提供更好、更具个性化的服务，就必须结合用户的需求，从这些多而杂的信息中找出真正有价值的商务信息. 面向商务信息领域的命名实体识别是实现商务信息自动提取的一项重要任务. 一般来说，命名实体识别任务是指识别文本中具有特定意义的实体，包括人名、地名、机构名、专有名词等，这也是目前国内外针对命名实体识别的主要研究方向. 然而在商务领域，产品信息的结构特点使产品命名实体的识别难度比一般命名实体要大，一般的命名实体识别技术并不能有效识别出产品实体. 笔者根据产品命名实体的特点，研究了一种三层的半监督学习算法，结合规则和统计的方法，更好地完成了面向商务信息的产品命名实体识别任务.

1 产品命名实体识别技术相关研究

命名实体识别主要有基于规则和基于统计 2 类方法. 基于规则的方法一般利用语言学专家所构造

收稿日期: 2012-10-09
基金项目: 国家科技支撑计划项目(2012BAH16F02); 国家自然科学基金项目(61003254)
作者简介: 黄诗琳(1984—)，女，博士生，E-mail: supercat@zju.edu.cn; 郑小林(1977—)，男，副教授.

的规则作为模板. 若建立的规则能很好地描述需要抽取的对象,则效果很好,但它们需要大量的领域知识背景和长期的投入. 基于统计的方法利用人工标注的语料进行训练学习,常用的统计机器学习方法都被应用到这一领域. 这类方法便于不同领域的移植,但对特征选择的要求较高. 在大多数情况下,更多是结合这 2 类方法来实现命名实体识别.

在商务信息处理领域,由于产品命名实体的结构特征多变,边界模糊,不适用一般命名实体识别方法. 根据命名实体的定义和已有的产品命名实体研究[1-3],一般认为产品命名实体包括品牌、系列、型号、种类和属性 5 个方面,但它们并非不可或缺. 一些品牌的产品没有系列名,而一些品牌独有的系列则可以省略品牌名. 另外,产品命名实体必须包含能指示具体品牌产品群体的信息,如"Sony 相机"是产品实体,"1 400 万像素相机"则不是.

随着商务智能化的迅速发展,目前国内外已经开始了一些针对产品命名实体识别的研究. 刘非凡、赵军、吕碧波等[1,4-5]提出了一种基于层级隐马尔可夫模型的产品命名实体识别方法,实现了汉语自由文本中产品命名实体识别和标注的原型系统. 罗芳等[6-7]引入本体特征,把条件随机场(CRF, conditional random field)算法应用到产品命名实体识别领域. 张朝胜等[8]同样使用了 CRF 算法,利用英文产品名特有的指示信息作为分类特征,结合手工构建的品牌词表进行建模.

为了降低基于统计的方法对人力标记语料的依赖,Li Xiao 等[9]以点击数据、产品元数据等额外的知识源作为手工标记的补充. 但是这种额外知识源一般很难取得. Usami 等[10]针对生物领域语料库的半自动获取建立了一个语义词典. 然而对于商务领域,产品实体形式多变,建立语义词典的难度较大. Irmak Utku 等[11]利用半结构化命名实体的特点,利用三层学习框架自举学习训练数据. 但是产品命名实体没有明显的结构特征,边界模糊,正则表达式难以建立.

三层结构的半监督 CRF 算法在第 1 层同时利用了产品命名实体各部分的结构特征和它们之间的关系特征,制定了一套专门的候选集提取策略. 然后在第 2 层通过一种优化的上下文相似度计算方法,找出与正例相似的候选词. 通过反复学习,得到足够的学习标记数据,并在第 3 层建立了一个隐条件随机场(HCRF, hidden conditional random field)模型,最终实现产品命名实体的识别.

2　三层的半监督 CRF 算法

与半结构命名实体不同,产品命名实体不具有明确结构特征,因此可以为它的 5 个部分采取不同的识别策略,反复自举学习出足够的训练数据,建立一个 HCRF 模型,把自举学习数据中无法正确识别的部分实体看作隐藏状态. 图 1 所示的三层的半监督算法整体框架.

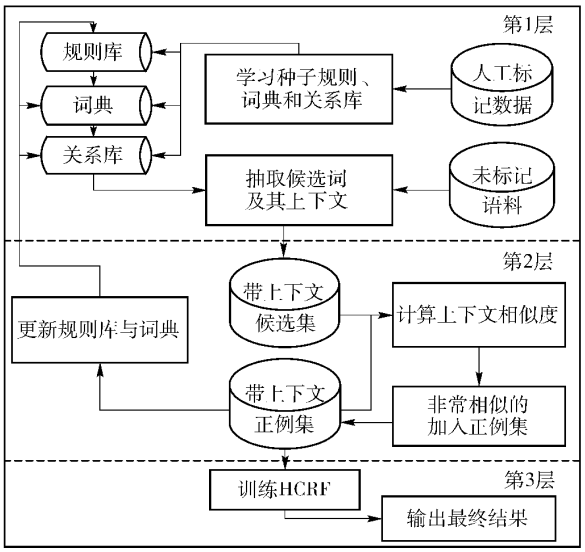

图 1　三层的半监督算法整体框架

2.1　第 1 层——候选集的选取

在一定产品领域内,品牌名、系列名和产品类别结构特征不明显,但是内容相对稳定,可以通过建立品牌词典、系列词典的方法来选取候选集合. 而对于产品型号,由于每天都有大量新产品推出,很难为产品型号建立词典. 但型号一般都是数字和字母的组合,因此可以利用规则对型号进行匹配. 对于产品属性,不同网站描述方式的差异使其表现形式变化繁多,尽管具体词汇不定,但某一类产品的属性种类是固定的,如相机具有颜色、焦距、大小等属性,并且前后可能出现一些指示词,因此可以根据这些特征提取出属性候选集.

产品命名实体的各部分分别具有其他命名实体的特征,只有以一定方式组合在一起才成为产品命名实体. 除了考虑它们各自的特征,还需要考虑它们的相互关系特征. 这可以通过分析手工标记数据中产品实体各部分的不同出现方式,建立一个关系规则库来描述关系特征. 只有同时满足实体自身特征和关系特征的词语才会被加入候选集. 同时,具

体产品名可以进行归并，虽然 Sony 和 Canon 是 2 个品牌，但在训练过程中都可以作为品牌候选词. 这种方式可以更大限度地利用手工标记数据.

2.2 第 2 层——上下文相似度比较

在确定候选集之后，需要识别出候选词是产品命名实体还是其他类别的命名实体，或者只是一个普通词. 与种子集上下文相似的候选词会被加入正例，直到获得足够正例.

候选词 c 的上下文 C 可以用向量空间模型来表示. 假设 C 具有 K 个特征词，那么 C 可以表示为 K 个特征词的集合 $W=\{w_1,w_2,\cdots,w_K\}$. w_k 在 C 中的重要程度用 $v(w_k,C)$ 表示，用 TF-IDF 方法计算. TF-IDF 是一种统计方法，TF 表示词频，词频越大则该词越重要:

$$\mathrm{tf}(w,C)=n_{w,C}\Big/\sum_j n_{w,C_j} \tag{1}$$

其中: $n_{w,C}$ 为特征词 w 在 C 中出现的次数，$\sum_j n_{w,C_j}$ 为归一化因子.

IDF 是逆向文档频率，表示词语的普遍重要性. 仅频繁出现在某一类别中的词很可能是这个类别的特征词. 然而，一般的 IDF 计算方式却会限制这些词出现高权重，因此可以把 IDF 公式优化为

$$\mathrm{idf}(w)=\lg\frac{Nn_{w,a}}{n+1}=\lg\frac{N}{1+n_{w,\bar{a}}+1/n_{w,a}} \tag{2}$$

其中: N 为上下文总数，n 为包含 w 的上下文数量，$n_{w,a}$ 为 w 在类别 A 中出现的次数，$n_{w,\bar{a}}$ 为 w 在其他类别中出现的次数. 这样 $\mathrm{idf}(w)$ 就随着 $n_{w,\bar{a}}/n_{w,a}$ 的增大而减少. 那么 w 的特征值 $v(w,C)$ 为

$$v(w,C)=\mathrm{tf}(w,C)\times\mathrm{idf}(w)=\frac{n_{w,C}}{\sum_j n_{w,C_j}}\times\lg\frac{Nn_{w,a}}{n+1} \tag{3}$$

相似度用 Tanimoto 系数计算. 记 $v(w_k,C)$ 为 $v_{C,k}$，S 是种子集，则 C 与上下文 $S_i\in S$ 的相似度为

$$\mathrm{sim}(V_C,V_{S_i})=\frac{\sum v_{C,k}v_{S_i,k}}{\sum v_{C,k}^2+\sum v_{S_i,k}^2-\sum v_{C,k}v_{S_i,k}} \tag{4}$$

C 与 S 的相似度 $\mathrm{sim}(C,S)$ 为 C 与所有 $S_i\in S$ 相似度的均值. 当 $\mathrm{sim}(C,S)$ 超过一定值时，候选词 c 作为产品命名实体加入 S 中.

2.3 第 3 层——HCRF 分类器

两层自举算法把相似度非常高的候选词放入正例，使一些相似度较低的实体无法被识别. 在 3 层的学习框架中，第 3 层把学习得到的内容作为输入，把未被识别出的实体作为隐藏状态，训练出一个 HCRF 分类器，输出最终的产品命名实体识别结果.

CRF 是一种判别式概率无向图模型，假设需要识别的文本序列为 $X=(x_1,x_2,\cdots,x_L)$，相应的标记序列为 $Y=(y_1,y_2,\cdots,y_L)$，那么对于产品命名实体识别任务，y_l 即为待识别实体的类别. 给定手工标记的训练数据 $T=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_T,Y_T)\}$ 和自举学习所得的训练数据 $T'=\{(X_1,Z_1),(X_2,Z_2),\cdots,(X_T,Z_T)\}$，由于只有高相似度的部分进行了标记，所以 T' 可能存在缺失. 若没有缺失，$y_t^i=z_t^i$; 若出现缺失，z_t^i 就是隐藏状态. 设学习所得的标签序列中隐藏状态集合为 $h=\{h_1,h_2,\cdots,h_M\}$，则标记序列的条件概率为[12]

$$p(y|x,\lambda)=\sum_h p(y,h|x,\lambda)=\frac{\mathrm{e}^{\Psi(y,h,x;\lambda)}}{\sum_{y',h}\mathrm{e}^{\Psi(y,h,x;\lambda)}} \tag{5}$$

其中 $\Psi(y,h,x;\lambda)$ 为一个势函数. 假设观察值的变动只与隐藏状态相关并且不包含边界，那么

$$\Psi(y,h,x;\lambda)=\sum_m\varphi(x_m)\lambda(h_m)+\sum_m\lambda(y,h_m) \tag{6}$$

其中 $\varphi(x_m)\lambda(h_m)$ 为观察序列和隐藏状态之间的共同特征参数. 根据最大熵模型，对数似然函数可以写成

$$E(\lambda)=\sum_{l=1}^{L}\lg\left(\frac{\sum_h\mathrm{e}^{\Psi(y,h,x_m;\lambda)}}{\sum_{y',h}\mathrm{e}^{\Psi(y,h,x_m;\lambda)}}\right)-\frac{1}{2\sigma^2}\sum_{l=1}^{L}\lambda_l^2 \tag{7}$$

其中 $\frac{1}{2\sigma^2}\sum_{l=1}^{L}\lambda_l^2$ 为避免过度学习的惩罚项.

这样，同时利用手工和学习取得的标记样本来训练 HCRF 分类器，可以最终识别出产品命名实体.

3 实验结果及分析

实验样本是随机收集的 700 个关于数码相机的新闻和评测网页文档，其中包含约 10 万字，共计 13 000多个产品实体. 从中选择 100 个网页文档作为种子集，100 个作为开放测试的测试集，并对种子集和测试集进行人工标记作为结果参考对比，其余 500 个页面作为无标记训练集. 评价标准采用 F-Measure，即准确率和召回率的几何平均值.

现有方法中，一般以基于 CRF 的有监督算法效果比较好，因此，把有监督 CRF 算法和本算法在同样的数据集上进行了对比测试. 图 2 和图 3 分别为手工标记页面数不同时封闭测试和开放测试的结果. 在手工标记数据较少的情况下，有监督 CRF 分类器不能很好地工作，而本算法可以利用无标记数据自举学习的训练数据作为补充，有效地保持了准确率. 而且，这种作用在开放测试中显得尤为明显.

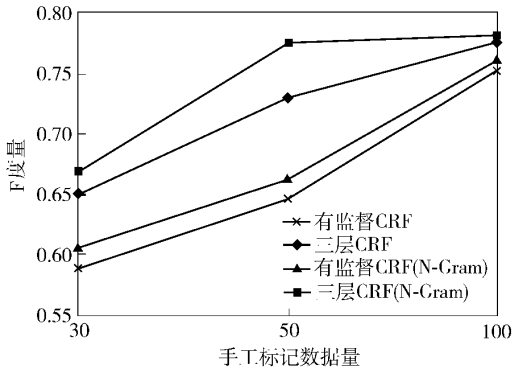

图 2　手工标记页面数不同时封闭测试结果

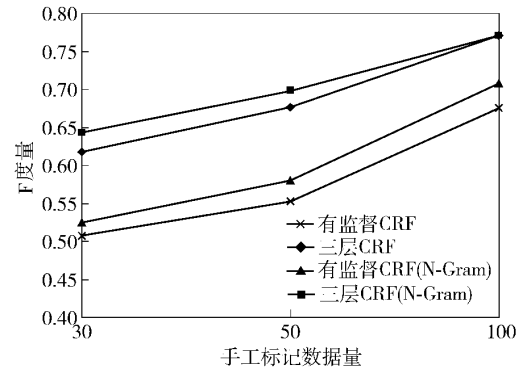

图 3　手工标记页面数不同时开放测试结果

表 1 所示为手工标记页面数都是 100 的情况下，无标记样本数量不同时的结果对比.

表 1　无标记数据集合大小对系统效果的影响

手工标记页面数	封闭测试		开放测试	
	三层 CRF	三层 CRF (N-Gram)	三层 CRF	三层 CRF (N-Gram)
100	0.773	0.779	0.746	0.754
200	0.775	0.778	0.751	0.758
500	0.776	0.782	0.770	0.771

在封闭测试中，准确率在无标记数据为 200 时反而比 100 时稍有下降. 可能的原因是在封闭测试下，自举算法错误学习的标记会对本来正确的标记造成干扰，而在开放测试中则不存在这样的问题.

综上所述，本算法在数码相机样本数据上的测试取得了较好的效果，但是相比一般命名实体识别还有一定的差距. 这主要是产品命名实体的边界本身比较模糊，在一定情况下还会与动词、助词等组成一个整体片段难以划分. 同时，数码相机领域命名实体的表达方式比一般命名实体更多样化，如相机镜头和相机本体可能单独作为一种产品，也可能两者组合成为单一产品“相机套机”，这种特殊性也加大了识别的困难.

4　结束语

产品命名实体识别是商务智能信息处理的重要基础. 三层结构的产品命名实体识别算法采用半监督的方式，结合规则和词典从无标记样本中获取高召回率的候选数据，把与种子集上下文相似的候选词加入正例，可以解决数据稀疏的问题. 自举学习中未能识别的实体作为隐藏状态输入，训练一个 HCRF 模型，有效地解决了产品命名实体识别的问题. 这种方法并不限定在数码相机或者电子产品领域，只要取得该领域的少量样本数据，也适用于其他类型的产品命名实体上. 但是该方法并未为产品的各项属性抽取出相应的值，在下一步的工作中将对此进行研究.

参考文献：

[1] Zhao Jun, Liu Feifan. Product named entity recognition in Chinese text [J]. Language Resources and Evaluation, 2008, 42(2): 197-217.

[2] Lertcheva N, Aroonmanakun W. A linguistic study of product names in Thai economic news [C] //2009 Eighth International Symposium on Natural Language Processing (SNLP'09). Bangkok: IEEE Press, 2009: 26-29.

[3] Lertcheva N, Aroonmanakun W. Product name identification and classification in Thai economic news [C] //Proceedings of the 2011 Named Entities Workshop (IJCNLP 2011). Chiang Mai: [s.n.], 2011: 58-64.

[4] 刘非凡，赵军，吕碧波，等. 面向商务信息抽取的产品命名实体识别研究 [J]. 中文信息学报，2006(1): 7-13.
Liu Feifan, Zhao Jun, Lü Bibo, et al. Study on product named entity recognition for business information extraction [J]. J Chin Inf Process, 2006(1): 7-13.

(下转第 54 页)

型,下一步工作将集中对蜂窝网络分布式干扰对齐算法进行深入研究,以实现干扰对齐技术在实际网络中的应用.

参考文献:

[1] Cadambe R, Jafar S. Interference alignment and degrees of freedom of the *K*-user interference channel [J]. IEEE Transactions on Information Theory, 2008, 54(8): 3425-3441.

[2] Gomadam K, Cadambe V, Jafar S, et al. A distributed numerical approach to interference alignment and applications to wireless interference networks [J]. IEEE Transactions on Information Theory, 2011, 57(6): 3309-3322.

[3] Peters W, Heath W. Cooperative algorithms for MIMO interference channels [J]. IEEE Transactions on Vehicular Technology, 2011, 60(1): 206-218.

[4] Shen Hui, Li Bin, Tao Meixia, et al. MSE-based transceiver designs for the MIMO interference channel [J]. IEEE Transactions on Wireless Communications, 2010, 9(11): 3480-3489.

[5] Monderer D, Shapley L. Potential games [J]. Games and Economic Behavior, 1996, 14(1): 124-143.

[6] Shi Changxin, Schmidt A, Berry A, et al. Distributed interference pricing for the MIMO interference channel [C] // 2009 IEEE International Conference on Communication (ICC2009). Dresden: [s. n.], 2009: 1-5.

(上接第23页)

[5] Liu Feifan, Zhao Jun, Lü Bibo, et al. Product named entity recognition based on hierarchical hidden markov model [C] // Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing. Jeju Island: [s. n.], 2005: 40-47.

[6] Luo Fang, Han Xiao, Chang Weili. Product named entity recognition using conditional random fields [C] // Business Intelligence and Financial Engineering (BIFE) 2011. Wuhan: IEEE Press, 2011: 86-89.

[7] Luo Fang, Qiu Qizhi, Xiong Qianxing. Introduction to the product-entity recognition task [C] //2011 3rd Symposium on Web Society (SWS). Port Elizabeth: IEEE Press, 2011: 122-126.

[8] 张朝胜,郭剑毅,线岩团,等. 基于条件随机场的英文产品命名实体识别 [J]. 计算机工程与科学, 2010, 32(6): 115-117.
Zhang Chaosheng, Guo Jianyi, Xian Yantuan, et al. Named entity recognition of the products with English based on conditional random fields [J]. Comput Eng Sci, 2010, 32(6): 115-117.

[9] Li Xiao, Wang Yeyi, Acero A. Extracting structured information from user queries with semi-supervised conditional random fields [C] // Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston: ACM, 2009: 572-579.

[10] Usami Yu, Cho Hancheol, Okazaki Naoaki, et al. Automatic acquisition of huge training data for bio-medical named entity recognition [C] // Proceedings of the 2011 Workshop on Biomedical Natural Language Processing. Portland: [s. n.], 2011: 65-73.

[11] Utku I, Reiner K. A scalable machine-learning approach for semi-structured named entity recognition [C] // Proceedings of the 19th International Conference on WWW. Raleigh: ACM, 2010: 461-470.

[12] Quattoni A, Wang Sybor, Morency L, et al. Hidden conditional random fields [J]. IEEE Trans Pattern Anal Mach Intell, 2007, 29(10): 1848-1852.